To: Unicode Technical Committee

L2/08-029

From: Peter Edberg, Apple Inc.

Date: Jan. 18, 2008

Re: UAX #29 Sentence Break SContinue characters

Document L2/07-399 proposed an enhancement to UAX #29 to improve sentence break for situations in which an ambiguous full stop after an abbreviation is followed by “continuation” punctuation (comma, colon, semicolon, em-dash) that removes the ambiguity, indicating that there is not a sentence break opportunity following that full stop. The enhancement included the addition of a new sentence-break character class SContinue and the modification of rule SB8a to include the SContinue class.

This enhancement was approved at UTC #113, and was incorporated in draft 3 of the proposed revision 12 to UAX #29 currently posted at <http://www.unicode.org/reports/tr29/tr29-12.html>. I was given action item 113-A26 to refine the list of characters that should be in the SContinue class (starting from the list in the original proposal), and forward the updated list to Mark Davis. Ken Whistler suggested that I use the collation data source to find other characters that are similar to comma, colon, semicolon, or em-dash.

In preparing the updated list, I noticed one problem: Semicolon can be a sentence-final character in Greek (indicating question mark). There is a script-specific character U+037E GREEK QUESTION MARK for this, but The Unicode Standard 5.0 indicates that U+003B SEMICOLON is preferred for this purpose. For now, rather than making the rules more complex to handle this particular case, I suggest not including semicolon-like characters in the SContinue class, in order to avoid breaking sentence-break behavior that currently works for Greek in order to fix one particular case of sentence-break behavior that does not currently work for Latin script. This means that sentence break will still find an incorrect break opportunity after “U.S.” in the following example from L2/07-399:

- The three groupings are Canada and the U.S.; France, Germany, and Denmark; and Italy, Spain, and Greece.

I sent Mark an updated list of SContinue characters that adds more characters similar to comma, colon, and em-dash, but deletes semicolon and similar characters. This list has been incorporated in the newly-posted draft 5 of the proposed revision 12 to UAX #29, with the character changes flagged as an open issue.

Thus, the original proposal L2/07-399 is modified as follows:

1R. Fill out the entry for SContinue in Table 4 with the following list of characters (bold indicates characters added to the original list in L2/07-399, strikethrough indicates characters deleted from the original list in L2/07-399; order is not significant):

SContinue	Any of the following characters: U+002C COMMA U+3001 IDEOGRAPHIC COMMA U+FE10 PRESENTATION FORM FOR VERTICAL COMMA U+FE11 PRESENTATION FORM FOR VERTICAL IDEOGRAPHIC COMMA U+FF0C FULLWIDTH COMMA **U+FE50 SMALL COMMA** **U+FF64 HALFWIDTH IDEOGRAPHIC COMMA** **U+FE51 SMALL IDEOGRAPHIC COMMA** **U+055D ARMENIAN COMMA** **U+060C ARABIC COMMA** **U+060D ARABIC DATE SEPARATOR** **U+07F8 NKO COMMA** **U+1802 MONGOLIAN COMMA** **U+1808 MONGOLIAN MANCHU COMMA** U+003A COLON U+FE13 PRESENTATION FORM FOR VERTICAL COLON U+FF1A FULLWIDTH COLON **U+FE55 SMALL COLON** U+2014 EM DASH U+FE31 PRESENTATION FORM FOR VERTICAL EM DASH U+002D HYPHEN-MINUS U+FF0D FULLWIDTH HYPHEN-MINUS **U+2013 EN DASH** **U+FE32 PRESENTATION FORM FOR VERTICAL EN DASH** **U+FE58 SMALL EM DASH** **U+FE63 SMALL HYPHEN-MINUS** ~~U+003B SEMICOLON~~ ~~U+FE14 PRESENTATION FORM FOR VERTICAL SEMICOLON~~ ~~U+FF1B FULLWIDTH SEMICOLON~~

3R. The text before SB6 needs to be further revised to remove the reference to semicolon (underline indicates new text, strikethrough indicates deleted text)**:**

Do not break after ambiguous terminators like period, if they are immediately followed by a number or lowercase letter, if they are between uppercase letters, if the first following letter (optionally after certain punctuation) is lowercase, or if they are followed by "continuation" punctuation such as comma <u>or</u> colon~~, or semicolon~~. For example, a period may be an abbreviation or numeric period, and thus may not mark the end of a sentence.